Ensemble-based Prototype-Augmented Multimodal Fusion for Ambivalence/Hesitancy Recognition

Elena Ryumina, Alexandr Axyonov, Dmitry Sysoev, Timur Abdulkadirov, Kirill Almetov, Yulia Morozova, Dmitry Ryumin

ryumina.e@iias.spb.su, axyonov.a@iias.spb.su, dsysoev@hse.ru, tnabdulkadirov@edu.hse.ru, koalmetov@edu.hse.ru, yuvmorozova_1@edu.hse.ru, daryumin@hse.ru

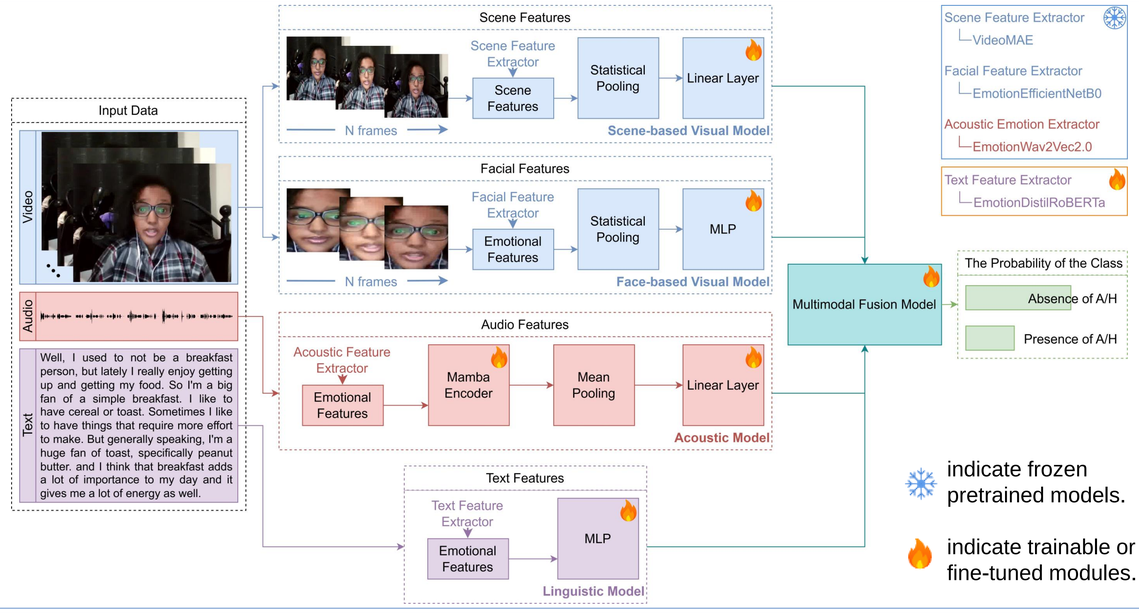
Introduction and Motivation

- A/H reflects subtle decision uncertainty, resistance, or fluctuating motivation, often expressed through cross-modal inconsistency between what a person says, how they speak, and how they appear.
- **Problem:** Current A/H recognition approaches are largely text-oriented, limiting their ability to model cross-modal inconsistencies between verbal, acoustic, facial, and scene cues.
- **Goal:** improving A/H recognition accuracy through multimodal fusion of scene, face, audio, and text cues while preserving modality-specific information

Methodology Overview

- Four dedicated unimodal branches are trained independently: VideoMAE scene model, EmotionEfficientNetB0 face model, EmotionWav2Vec2.0+Mamba acoustic model, and EmotionDistilRoBERTa linguistic model.
- Each branch produces a compact video-level embedding; embeddings are projected into a shared latent space and treated as modality tokens.
- Transformer-based multimodal fusion models inter-modality dependencies.
- A prototype-augmented variant adds learnable class-specific prototypes as an auxiliary loss to structure the fused representation.

Proposed Pipeline



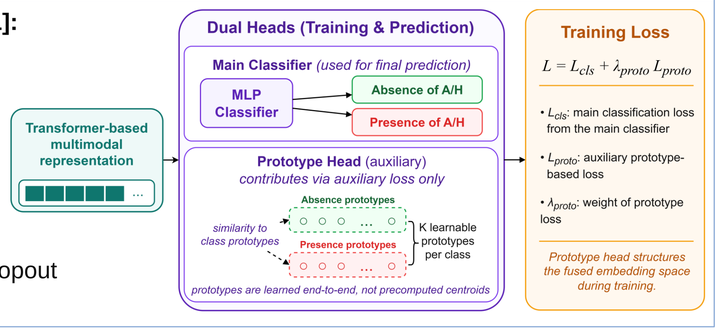
Research Corpus and Training Setup

Behavioural Ambivalence/Hesitancy corpus [1]:
1,427 videos · 300 participants · 10.60 hours

Fusion search:
Optuna · 5 seeds · stability-oriented selection

Final system:
5 prototype-augmented fusion models ensemble

Best multimodal fusion model:
latent size 128, 6 Transformer layers, 4 heads, dropout 0.45, 16 prototypes per class, temperature τ=0.3



Experimental Results on ABAW'26 BAH Challenge

Configuration	Modalities	Public devel.	Public test	Average	Private test
Face: EmotionEfficientNetB0 + statistics	F	65.29	60.05	62.67	—
Scene: VideoMAE	S	61.71	62.21	61.96	—
Audio: EmotionWav2Vec2.0 + Mamba	A	67.20	70.87	69.03	—
Text: fine-tuned EmotionDistilRoBERTa	T	68.54	71.49	70.02	—
MFM without prototype head	A+F+S+T	85.38	79.94	82.66	68.32
MFM with prototype head	A+F+S+T	83.79	82.72	**83.25**	**65.21**
Ensemble of five MFMs without prototype head	A+F+S+T	81.94	80.64	81.29	70.17
Ensemble of five MFMs with prototype head	A+F+S+T	83.00	80.77	**81.89**	**71.42**

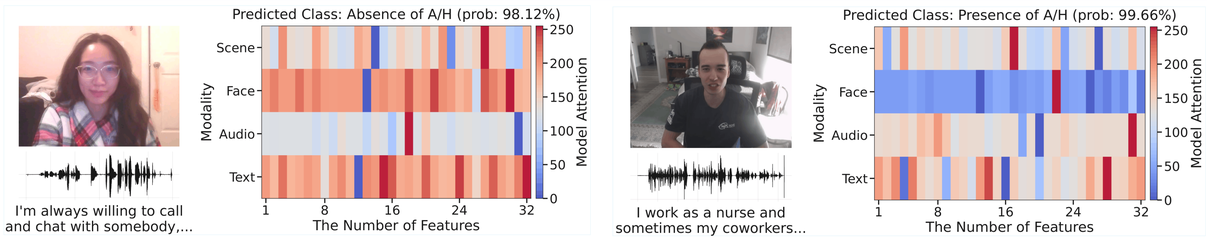
Approach	Modalities	Private test
Pereira et al. [2]	A+F+T	72.66
Bekhouche et al. [3]	A+F+T	71.51
Ours [4]	A+F+S+T	**71.42**
Tang et al. [4]	M-LLM	67.48
Baseline	M-LLM	34.28

MF1 performance. MFM means Multimodal Fusion Model

Key findings

- Multimodal fusion clearly surpasses all unimodal baselines.
- Best average performance on public development and test subsets: 83.25% MF1 with prototype-augmented fusion.
- Best private test score: 71.42% MF1 with five-model prototype ensemble.

Modality-wise Attention



- Example without A/H: the model assigns 98.12% confidence and relies primarily on face and text cues.
- Example with A/H: the model assigns 99.66% confidence and shifts attention toward audio and text.
- The fusion model adaptively reweights modalities per sample, which is important when relevant cues are inconsistent across face, speech, language, and scene context.

Discussion and Conclusions

- Text is the strongest unimodal source (70.02% average MF1), but scene and audio provide complementary evidence.
- The ablation study shows that scene+text is the strongest two-modality pair (80.39% average MF1).
- Prototype-based auxiliary supervision helps structure the fused latent space; ensembling is crucial for private-test robustness.
- Overall, prototype-augmented multimodal fusion is an effective strategy for A/H recognition in unconstrained videos.

References

[1] Gonzalez-González et al., BAH Dataset for Ambivalence/Hesitancy Recognition in Videos for Digital Behavioural Change, ICLR 2026.
[2] Pereira et al., Behavioral recognition optimized through heterogeneous ensemble regularization for ambivalence and hesitancy. ArXiv 2026.
[3] Bekhouche et al., Conflict-aware multimodal fusion for ambivalence and hesitancy recognition, ArXiv 2026.
[4] Ryumina et al., Team LEYA in 10th ABAW Competition: Multimodal Ambivalence/Hesitancy Recognition Approach, ArXiv 2026.
[5] Tang et al., Nuance demotion recognition based on a segment-based MLLM frame-work leveraging Qwen3-Omni for FH detection, ArXiv 2026.